Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 5, *pp.*

Conditional transformation models

Torsten Hothorn,

Ludwig-Maximilians-Universität München, Germany, and Universität Zürich, Switzerland

Thomas Kneib

Georg-August-Universität Göttingen, Germany

and Peter Bühlmann

Eidgenössische Technische Hochschule, Zürich, Switzerland

[Received January 2012. Revised November 2012]

Summary. The ultimate goal of regression analysis is to obtain information about the conditional distribution of a response given a set of explanatory variables. This goal is, however, seldom achieved because most established regression models estimate only the conditional mean as a function of the explanatory variables and assume that higher moments are not affected by the regressors. The underlying reason for such a restriction is the assumption of additivity of signal and noise. We propose to relax this common assumption in the framework of transformation models. The novel class of semiparametric regression models proposed herein allows transformation functions to depend on explanatory variables. These transformation functions are estimated by regularized optimization of scoring rules for probabilistic forecasts, e.g. the continuous ranked probability score. The corresponding estimated conditional distribution functions are consistent. Conditional transformation models are potentially useful for describing possible heteroscedasticity, comparing spatially varying distributions, identifying extreme events, deriving prediction intervals and selecting variables beyond mean regression effects. An empirical investigation based on a heteroscedastic varying-coefficient simulation model demonstrates that semiparametric estimation of conditional distribution functions can be more beneficial than kernel-based non-parametric approaches or parametric generalized additive models for location, scale and shape.

Keywords: Boosting; Conditional distribution function; Conditional quantile function; Continuous ranked probability score; Prediction intervals; Structured additive regression

1. Introduction

One of the famous 'Top 10 reasons to become a statistician' is that statisticians are 'mean lovers' (Friedman *et al.*, 2002), referring of course to our obsession with means. Whenever a distribution is too complex to think or expound on, we focus on the mean as a single real number describing the centre of the distribution and we block out other characteristics such as variance, skewness and kurtosis. Our willingness to simplify distributions this way is most apparent when we deal with many distributions at a time, as in a regression setting where we describe the conditional

Address for correspondence: Torsten Hothorn, Institut für Sozial- und Präventivmedizin, Abteilung Biostatistik, Universität Zürich, Hirschengraben 84, CH-8001 Zürich, Switzerland.
E-mail: Torsten.Hothorn@R-project.org

 1369–7412/13/75000

distribution $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ of a response $Y \in \mathbb{R}$ given different configurations of explanatory variables $\mathbf{X} = \mathbf{x} \in \chi$. Many regression models focus on the conditional mean $\mathbb{E}(Y|\mathbf{X}=\mathbf{x})$ and treat higher moments of the conditional distribution $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ as fixed or nuisance parameters that must not depend on the explanatory variables. As a consequence, model inference crucially relies on assumptions such as homoscedasticity and symmetry. Information on the scale of the response, e.g. prediction intervals, derived from such models also depends on these assumptions. Here, we propose a new class of conditional transformation models that allow the conditional distribution function $\mathbb{P}(Y \leqslant v|\mathbf{X}=\mathbf{x})$ to be estimated directly and semiparametrically under quite weak assumptions. Before we introduce this class of models, we shall attempt to set a scene of contemporary regression in the light of Gilchrist (2008) and to place the major players on this stage.

Let $Y_{\mathbf{x}} = (Y|\mathbf{X}=\mathbf{x}) \sim \mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ denote the conditional distribution of response Y given explanatory variables $\mathbf{X}=\mathbf{x}$. We assume that $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$ is dominated by some measure μ and has the conditional distribution function $\mathbb{P}(Y \leqslant v|\mathbf{X}=\mathbf{x})$. A regression model describes the distribution $\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}$, or certain characteristics of it, as a function of the explanatory variables $\mathbf{x}$. We estimate such models on the basis of samples of pairs of random variables $(Y, \mathbf{X})$ from the joint distribution $\mathbb{P}_{Y,\mathbf{X}}$. It is convenient to assume that a regression model consists of signal and noise, i.e. a deterministic part and an error term. In what follows, we denote the error term by $Q(U)$, where $U \sim \mathcal{U}[0,1]$ is a uniform random variable independent of $\mathbf{X}$ and $Q : \mathbb{R} \to \mathbb{R}$ is the quantile function of an absolutely continuous distribution.

Apart from non-parametric kernel estimators of the conditional distribution function (Hall *et al.*, 1999; Hall and Müller, 2003; Li and Racine, 2008), there are two common ways to model the influence of the explanatory variables $\mathbf{x}$ on the response $Y_{\mathbf{x}}$:

$$Y_{\mathbf{x}} = r\{Q(U)|\mathbf{x}\} \qquad \text{'mean or quantile regression models'} \tag{1}$$

and

$$h(Y_{\mathbf{x}}|\mathbf{x}) = Q(U) \qquad \text{'transformation models'}. \tag{2}$$

For each $\mathbf{x} \in \chi$, the regression function $r(\cdot|\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ transforms the error term $Q(U)$ in a monotone increasing way. The inverse regression function $h(\cdot|\mathbf{x}) = r^{-1}(\cdot|\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ is also monotone increasing. Because h transforms the response, it is known as a transformation function and models in the form of equation (2) are called transformation models.

A major assumption underlying almost all regression models of class (1) is that the regression function r is the sum of the deterministic part $r_{\mathbf{x}} : \chi \to \mathbb{R}$, which depends on the explanatory variables, and the error term:

$$r\{Q(U)|\mathbf{x}\} = r_{\mathbf{x}}(\mathbf{x}) + Q(U).$$

When $\mathbb{E}\{Q(U)\} = 0$, we obtain, $r_{\mathbf{x}}(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X}=\mathbf{x})$, e.g. linear or additive models depending on the functional form of $r_{\mathbf{x}}$. Model inference is commonly based on the normal error assumption, i.e. $Q(U) = \sigma\Phi^{-1}(U)$, where $\sigma > 0$ is a scale parameter and $\Phi^{-1}(U) \sim \mathcal{N}(0,1)$. A novel semiparametric approach is extended generalized additive models, where additive functions of the explanatory variables describe location, scale and shape (generalized additive models for location, scale and shape (GAMLSSs)) of a certain parametric conditional distribution of the response given the explanatory variables (Rigby and Stasinopoulos, 2005). If the assumption of a certain parametric form of the conditional distribution is questionable, $r_{\mathbf{x}}$ describes the τ-quantile of $Y_{\mathbf{x}}$ when the quantile function Q is such that $Q(\tau) = 0$ for some $\tau \in (0,1)$. This leads us to quantile regression (Koenker, 2005). Estimating the complete conditional quantile function is less straightforward since we must fit separate models for a grid of probabili-

ties τ, and the resulting regression quantiles may cross. Solutions to this problem can be obtained by combining all quantile fits in one joint model based on, for example, location–scale models (He, 1997) or quantile sheets (Schnable and Eilers, 2012), or by monotonizing the estimated quantile curves by using non-decreasing rearrangements (Dette and Volgushev, 2008).

For transformation models (2), additivity is assumed on the scale of the inverse regression function h:

$$h(Y_{\mathbf{x}}|\mathbf{x}) = h_Y(Y_{\mathbf{x}}) + h_{\mathbf{x}}(\mathbf{x}).$$

When $\mathbb{E}\{Q(U)\} = 0$, we obtain $-h_{\mathbf{x}}(\mathbf{x}) = \mathbb{E}\{h_Y(Y_{\mathbf{x}})\} = \mathbb{E}\{h_Y(Y)|\mathbf{X} = \mathbf{x}\}$. The monotone transformation function $h_Y : \mathbb{R} \to \mathbb{R}$ does not depend on $\mathbf{x}$ and might be known in advance (Box–Cox transformation models with fixed parameters, accelerated failure time models) or is commonly treated as a nuisance parameter (Cox model; proportional odds model). One is usually interested in estimating the function $h_{\mathbf{x}} : \chi \to \mathbb{R}$, which describes the conditional mean of the *transformed* response $h_Y(Y_{\mathbf{x}})$. The class of transformation models is rich and very actively researched, most prominently in literature on the analysis of survival data. For example, in the Cox additive model, $h_Y(Y_{\mathbf{x}}) = \log\{\Lambda(Y_{\mathbf{x}})\}$ is based on the unspecified integrated baseline hazard function Λ, $h_{\mathbf{x}}(\mathbf{x}) = \Sigma_{j=1}^{J} h_{\mathbf{x},j}(\mathbf{x})$ is the sum of J smooth terms depending on the explanatory variables and $Q(U) = -\log\{-\log(U)\}$ is the quantile function of the extreme value distribution. Doksum and Gasko (1990) discussed the flexibility of this class of models, and Cheng *et al.* (1995) introduced a generic algorithm for linear transformation model estimation, i.e. with $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$, treating h_Y as a nuisance.

In recent years, transformation models have been extended in two directions. In the first direction, more flexible forms for the conditional mean function $h_{\mathbf{x}}$ have been introduced, e.g. the partially linear transformation model $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}(0, \boldsymbol{\alpha})^{\mathrm{T}} + h_{\mathrm{smooth}}(x_1)$ (where h_{smooth} is a smooth function of the first variable x_1; Lu and Zhang (2010)), the varying-coefficient model $h_{\mathbf{x}}(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}(0, 0, \boldsymbol{\alpha})^{\mathrm{T}} + h_{\mathrm{smooth}}(x_1)x_2$ (Chen and Tong, 2010), random-effects models (Zeng *et al.*, 2005) and various approaches to additive transformation and accelerated failure time models, such as the boosting approaches by Lu and Li (2008) and Schmid and Hothorn (2008). In the second direction, many researchers have considered algorithms that estimate h_Y and (partially) linear functions $h_{\mathbf{x}}$ simultaneously, usually by a spline expansion of h_Y (e.g. Shen (1998) and Cheng and Wang (2011)), as an alternative to the common practice of estimating h_Y *post hoc* by some non-parametric procedure such as the Breslow estimator.

Although the transformation function h_Y is typically treated as an infinite dimensional nuisance parameter, it is important to note that h_Y contains information about higher moments of $Y_{\mathbf{x}}$, most importantly variance and skewness. Simultaneous estimation of h_Y and $h_{\mathbf{x}}$ is therefore extremely attractive because we can obtain information about the mean and higher moments of the transformed response at the same time. However, owing to the decomposition of the regression function r or the transformation function h into both a deterministic part depending on the explanatory variables ($r_{\mathbf{x}}$ or $h_{\mathbf{x}}$) and a random part depending on the response (h_Y) or error term ($Q(U)$), higher moments of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ must not depend on the explanatory variables in mean regression and transformation models. As a consequence, the corresponding models cannot capture heteroscedasticity or skewness induced by certain configurations of the explanatory variables. Therefore, we cannot detect these potentially interesting patterns, and our models will perform poorly when probability forecasts, prediction intervals or other functionals of the conditional distribution are of special interest.

Recently, Wu *et al.* (2010) proposed a novel transformation model for longitudinal data that partially addresses this issue. For responses and explanatory variables $\mathbf{X}(t)$ observed at

time t, the model assumes that $h(Y_{\mathbf{x}}|t, \mathbf{x}) = h_Y(Y_{\mathbf{x}}|t) + \mathbf{x}(t)^{\mathrm{T}}\boldsymbol{\alpha}(t)$. Here, the transformation h_Y is conditional on time, and higher moments may vary with time. However, since h_Y does not depend on the explanatory variables $\mathbf{x}$, these higher moments may not vary with one or more of the explanatory variables. In the context of longitudinal data with functional explanatory variables, Chen and Müller (2012) considered a similar model, where the regression coefficients for functional principal components may depend on time t and the response $Y_{\mathbf{x}}$. Our contribution is a class of transformation models where the transformation function is conditional on the explanatory variables in the sense that the transformation function, and therefore higher moments of the conditional distribution of the response, may depend on potentially all explanatory variables. As a consequence, the models that are suggested here can deal with heteroscedasticity and skewness that can be regressed on the explanatory variables, and we shall show that reliable estimates of the complete conditional distribution function and functionals thereof can be obtained.

We shall introduce these 'conditional transformation models' (Section 2), discuss the underlying model assumptions and embed the estimation problem in the empirical risk minimization framework (Section 3). For simplicity, we restrict ourselves to continuous responses Y that have been observed without censoring. We present a computationally efficient algorithm for fitting the models in Section 4 and study the asymptotic properties of the estimated conditional distribution functions in Section 5. The practical benefits of modelling the influence of explanatory variables on the variance and higher moments of the response distribution are demonstrated in Section 6 with a special emphasis on distributional characteristics of childhood nutrition in India and on prediction intervals for birth weights of small fetuses. Finally, we use a heteroscedastic varying-coefficient simulation model to evaluate the empirical performance of the algorithm proposed and compare the quality of conditional distribution functions estimated by a conditional transformation model and established parametric and non-parametric procedures in Section 7.

2. Conditional transformation models

An attractive feature of transformation models is their close connection to the conditional distribution function. With the transformation function $h(Y_{\mathbf{x}}|\mathbf{x}) = Q(U)$, we can evaluate the conditional distribution function of response Y given the explanatory variables $\mathbf{x}$ via

$$\mathbb{P}(Y \leqslant v|\mathbf{X} = \mathbf{x}) = \mathbb{P}\{h(Y|\mathbf{x}) \leqslant h(v|\mathbf{x})\} = F\{h(v|\mathbf{x})\}$$

with absolute continuous distribution function $F = Q^{-1}$. For additive transformation functions $h = h_Y + h_{\mathbf{x}}$, the conditional distribution function reads $F\{h(v|\mathbf{x})\} = F\{h_Y(v) + h_{\mathbf{x}}(\mathbf{x})\}$, i.e. the distribution is evaluated for a transformed and shifted version of Y. Higher moments depend only on the transformation h_Y and thus cannot be influenced by the explanatory variables. Consequently, we must avoid the additivity in the model $h = h_Y + h_{\mathbf{x}}$ to allow the explanatory variables to impact also higher moments. We therefore suggest a novel transformation model based on an alternative additive decomposition of the transformation function h into J partial transformation functions for all $\mathbf{x} \in \chi$:

$$h(v|\mathbf{x}) = \sum_{j=1}^{J} h_j(v|\mathbf{x}), \tag{3}$$

where $h(v|\mathbf{x})$ is the monotone transformation function of v. In this model, the transformation function $h(Y_{\mathbf{x}}|\mathbf{x})$ and the partial transformation functions $h_j(\cdot|\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ are conditional on $\mathbf{x}$ in the sense that not only the mean of $Y_{\mathbf{x}}$ depends on the explanatory variables. For this reason,

we coin models of the form (3) *conditional transformation models*. Clearly, model (3) imposes an assumption, namely additivity of the conditional distribution function on the scale of the quantile function Q, i.e.

$$Q\{\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})\} = \sum_{j=1}^{J} h_j(\upsilon|\mathbf{x}).$$

It should be noted that here we assume additivity of the transformation function h and not additivity on the scale of the regression function r as is common for additive mean or quantile regression models (1). Furthermore, monotonicity of h_j is sufficient but not necessary for h being monotone. Of course, we have to make further assumptions on h_j to obtain reasonable models, but these assumptions are problem specific, and we shall therefore postpone these issues until Section 6. To ensure identifiability, we assume without loss of generality that the partial transformation functions are centred around zero, $\mathbb{E}_Y[\mathbb{E}_{\mathbf{X}}\{h_j(Y|\mathbf{X})\}]=0$ for all $j=1,\ldots,J$ for non-systematic error terms ($\{\mathbb{E}(Q(U)\}=0$).

3. Estimating conditional transformation models

The estimation of conditional distribution functions can be reformulated as a mean regression problem since $\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) = \mathbb{E}\{I(Y \leqslant \upsilon)|\mathbf{X}=\mathbf{x}\}$ for the binary event $Y \leqslant \upsilon$; this connection is widely used (e.g. by Hall and Müller (2003) and Chen and Müller (2012)). Similarly to the approach of fitting multiple quantile regression models to obtain an estimate of the conditional quantile function, one could estimate the models $\mathbb{E}\{I(Y \leqslant \upsilon)|\mathbf{X}=\mathbf{x}\}$ for a grid of υ values separately. However, we instead suggest estimating conditional transformation models by the application of an integrated loss function that allows the whole conditional distribution function to be obtained in one step.

Let ρ denote a function of measuring the loss of the probability $F\{h(\upsilon|\mathbf{X})\}$ for the binary event $Y \leqslant \upsilon$. One candidate loss function is

$$\begin{aligned}\rho_{\text{bin}}\{(Y \leqslant \upsilon, \mathbf{X}), h(\upsilon|\mathbf{X})\} := -(&I(Y \leqslant \upsilon)\log[F\{h(\upsilon|\mathbf{X})\}] \\ &+ \{1 - I(Y \leqslant \upsilon)\}\log[1 - F\{h(\upsilon|\mathbf{X})\}]) \geqslant 0,\end{aligned}$$

the negative log-likelihood of the binomial model $(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) \sim B[1, F\{h(\upsilon|\mathbf{x})\}]$ for the binary event $Y \leqslant \upsilon$ with link function $Q = F^{-1}$. Alternatively, we may consider the squared or absolute error losses

$$\begin{aligned}\rho_{\text{sqe}}\{(Y \leqslant \upsilon, \mathbf{X}), h(\upsilon|\mathbf{X})\} &:= 0.5|I(Y \leqslant \upsilon) - F\{h(\upsilon|\mathbf{X})\}|^2 \geqslant 0, \\ \rho_{\text{abe}}\{(Y \leqslant \upsilon, \mathbf{X}), h(\upsilon|\mathbf{X})\} &:= |I(Y \leqslant \upsilon) - F\{h(\upsilon|\mathbf{X})\}| \geqslant 0.\end{aligned}$$

The squared error loss ρ_{sqe} is also known as the Brier score and the absolute loss ρ_{abe} has been applied for assessing survival probabilities in the Cox model by Schemper and Henderson (2000). We define the loss function l for estimating conditional transformation models as integrated loss ρ with respect to a measure μ dominating the conditional distribution $\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})$:

$$l\{(Y, \mathbf{X}), h\} := \int \rho\{(Y \leqslant \upsilon, \mathbf{X}), h(\upsilon|\mathbf{X})\}\,\mathrm{d}\mu(\upsilon) \geqslant 0.$$

In the context of scoring rules, the loss l based on ρ_{sqe} is known as the continuous ranked probability score or integrated Brier score and is a proper scoring rule for assessing the quality of probabilistic or distributional forecasts (see Gneiting and Raftery (2007) for an overview). It seems natural to apply these scores as loss functions for model estimation, but we are aware

of only the work of Gneiting *et al.* (2005), who directly optimized the continuous ranked probability score for estimating Gaussian predictive probability density functions for continuous weather variables. In the context of non-parametric or semiparametric estimation of conditional distribution functions, minimization of the empirical analogue of the risk function

$$\mathbb{E}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),h\}] = \int\int \rho\{(y \leqslant \upsilon, \mathbf{x}), h(\upsilon|\mathbf{x})\}\, \mathrm{d}\mu(\upsilon)\, \mathrm{d}\mathbb{P}_{Y,\mathbf{X}}(y,\mathbf{x}) \geqslant 0$$

for estimating conditional distribution functions has not yet been considered. Model estimation based on the risk $\mathbb{E}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),h\}]$ is reasonable because the corresponding optimization problem is convex and attains its minimum for the true conditional transformation function h. We summarize these facts in the following lemma, whose proof in given in Appendix A.

Lemma 1. The risk $\mathbb{E}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),h\}]$ is convex in h for convex losses ρ in h. The population minimizer of $\mathbb{E}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),h\}]$ for $\rho = \rho_{\text{bin}}$ and $\rho = \rho_{\text{sqe}}$ is $h(\upsilon|\mathbf{x}) = Q\{\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})\}$. For $\rho = \rho_{\text{abe}}$, the minimizer is

$$h(\upsilon|\mathbf{x}) = \begin{cases} -\infty, & \mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) \leqslant 0.5, \\ \infty, & \mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) > 0.5. \end{cases}$$

The corresponding empirical risk function defined by the data is

$$\hat{\mathbb{E}}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),f\}] = \int\int \rho\{(y \leqslant \upsilon, \mathbf{x}), h(\upsilon|\mathbf{x})\}\, \mathrm{d}\mu(\upsilon)\, \mathrm{d}\hat{\mathbb{P}}_{Y,\mathbf{X}}(y,\mathbf{x}) \geqslant 0.$$

On the basis of an independent and identically distributed random sample $(Y_i, \mathbf{X}_i) \sim \mathbb{P}_{Y,\mathbf{X}}$, $i = 1, \ldots, N$, of N observations from the joint distribution of response and explanatory variables, we define $\hat{\mathbb{P}}_{Y,\mathbf{X}}$ as the distribution putting mass $w_i > 0$ on observation i ($w_i \equiv N^{-1}$ for the empirical distribution). For computational convenience, we also approximate the measure μ by the discrete uniform measure $\hat{\mu}$, which puts mass n^{-1} on each element of the equidistant grid $\upsilon_1 < \ldots < \upsilon_n \in \mathbb{R}$ over the response space. The number of grid points n must be sufficiently large to approximate the integral closely. The empirical risk is then

$$\hat{\mathbb{E}}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),h\}] = n^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} w_i\, \rho\{(Y_i \leqslant \upsilon_\imath, \mathbf{X}_i), h(\upsilon_\imath|\mathbf{X}_i)\}. \tag{4}$$

This risk is the weighted empirical risk for loss function ρ evaluated at the observations $(Y_i \leqslant \upsilon_\imath, \mathbf{X}_i)$ for $i = 1, \ldots, N$ and $\imath = 1, \ldots, n$. Consequently, we can apply algorithms for fitting generalized additive models to the binary responses $Y_i \leqslant \upsilon_\imath$ under loss ρ for estimating model (3). Although this seems to be quite straightforward, there are two issues to consider. First, simply expanding the observations over the grid $\upsilon_1 < \ldots < \upsilon_n$ increases the computational complexity by n, which, even for moderately large sample sizes N, renders computing and storage rather burdensome. Second, unconstrained minimization of the empirical risk, i.e. no smoothness of h in its first argument and h being independent of the conditioning $\mathbf{x}$, leads to estimates $F\{\hat{h}(\upsilon|\mathbf{x})\} = \hat{\mathbb{P}}(Y \leqslant \upsilon) = N^{-1} \Sigma_{i=1}^{N} I(Y_i \leqslant \upsilon)$, i.e. the empirical cumulative distribution function of Y for ρ_{bin} and ρ_{sqe} with $w_i = N^{-1}$. For ρ_{abe}, the empirical risk is minimized by $F\{\hat{h}(\upsilon|\mathbf{x})\} = 0$ for all υ with $\hat{\mathbb{P}}(Y \leqslant \upsilon) < 0.5$ and otherwise by $F\{\hat{h}(\upsilon|\mathbf{x})\} = 1$.

Therefore, careful regularization is absolutely necessary to obtain reasonable models that lead to smooth conditional distribution functions (i.e. smoothing in the Y-direction) and that are similar for similar configurations of the explanatory variables (i.e. smoothing in the $\mathbf{X}$-direction). Instead of adding a direct penalization term to the empirical risk, we propose in the next section

a boosting algorithm for empirical risk minimization that indirectly controls the functional form and complexity of the estimate $\hat{h}$.

4. Boosting conditional transformation models

We propose to fit conditional transformation models (3) by a variant of componentwise boosting for minimizing equation (4) with penalization. In this class of algorithms, regularization is achieved indirectly via the application of penalized base learners, and the complexity of the whole model is controlled by the number of boosting iterations. We refer the reader to Bühlmann and Hothorn (2007) for a detailed introduction to componentwise boosting.

For conditional transformation models, we parameterize the partial transformation functions for all $j = 1, \ldots, J$ as

$$h_j(\upsilon|\mathbf{x}) = \{\mathbf{b}_j(\mathbf{x})^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon)^{\mathrm{T}}\}\boldsymbol{\gamma}_j \in \mathbb{R}, \qquad \boldsymbol{\gamma}_j \in \mathbb{R}^{K_j K_0}, \tag{5}$$

where $\mathbf{b}_j(\mathbf{x})^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon)^{\mathrm{T}}$ denotes the tensor product of two sets of basis functions $\mathbf{b}_j : \chi \to \mathbb{R}^{K_j}$ and $\mathbf{b}_0 : \mathbb{R} \to \mathbb{R}^{K_0}$. Here, $\mathbf{b}_0$ is a basis along the grid of υ-values that determines the functional form of the response transformation. The basis $\mathbf{b}_j$ defines how this transformation may vary with certain aspects of the explanatory variables. The tensor product may be interpreted as a generalized interaction effect (which is further illustrated in Section 6). For each partial transformation function h_j, we typically want to obtain an estimate that is smooth in its first argument υ and smooth in the conditioning variable $\mathbf{x}$. Therefore, the bases are supplemented with appropriate, prespecified penalty matrices $\mathbf{P}_j \in \mathbb{R}^{K_j \times K_j}$ and $\mathbf{P}_0 \in \mathbb{R}^{K_0 \times K_0}$, inducing the penalty matrix $\mathbf{P}_{0j} = (\lambda_0 \mathbf{P}_j \otimes \mathbf{1}_{K_0} + \lambda_j \mathbf{1}_{K_j} \otimes \mathbf{P}_0)$ with smoothing parameters $\lambda_0 \geqslant 0$ and $\lambda_j \geqslant 0$ for the tensor product basis. The base learners corresponding to the partial transformation functions fitted to the negative gradients in each iteration of the boosting algorithm are then ridge-type linear models with penalty matrix $\mathbf{P}_{0j}$. In more detail, we apply the following algorithm for fitting conditional transformation models with transformation functions (5).

4.1. Algorithm: boosting for conditional transformation models

Step 1 (initialization): initialize $\boldsymbol{\gamma}_j^{[0]} \equiv 0$ for $j = 1, \ldots, J$, the step size $\nu \in (0, 1)$ and the smoothing parameters λ_j, $j = 0, \ldots, J$. Define the grid $\upsilon_1 \leqslant Y_{(1)} < \ldots < Y_{(N)} \leqslant \upsilon_n$. Set $m := 0$.
Step 2 (gradient): compute the negative gradient $U_{i\iota}$ for $\hat{h}_{i\iota}^{[m]} = \sum_{j=1}^{J} \{\mathbf{b}_j(\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon_\iota)^{\mathrm{T}}\} \boldsymbol{\gamma}_j^{[m]}$,

$$U_{i\iota} := - \frac{\partial}{\partial h} \rho\{(Y_i \leqslant \upsilon_\iota, \mathbf{X}_i), h\} \Big|_{h = \hat{h}_{i\iota}^{[m]}}.$$

Fit the base learners for $j = 1, \ldots, J$ with penalty matrix $\mathbf{P}_{0j}$:

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K_j K_0}}{\arg\min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} w_i [U_{i\iota} - \{\mathbf{b}_j(\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon_\iota)^{\mathrm{T}}\}\boldsymbol{\beta}]^2 + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{P}_{0j} \boldsymbol{\beta}. \tag{6}$$

Select the best base learner:

$$j^* = \underset{j = 1, \ldots, J}{\arg\min} \sum_{i=1}^{N} \sum_{\iota=1}^{n} w_i [U_{i\iota} - \{\mathbf{b}_j(\mathbf{X}_i)^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon_\iota)^{\mathrm{T}}\}\hat{\boldsymbol{\beta}}_j]^2.$$

Step 3 (update): update the parameters $\boldsymbol{\gamma}_{j^*}^{[m+1]} = \boldsymbol{\gamma}_{j^*}^{[m]} + \nu \hat{\boldsymbol{\beta}}_{j^*}$ and keep all other parameters fixed, i.e. $\boldsymbol{\gamma}_j^{[m+1]} = \boldsymbol{\gamma}_j^{[m]}$, $j \neq j^*$. Iterate steps 2 and 3.

Step 4 (stop): stop if $m = M$. Output the final model as a function of arbitrary $\upsilon \in \mathbb{R}$ and $\mathbf{x} \in \chi$:

$$\hat{\mathbb{P}}(Y \leqslant \upsilon | \mathbf{X} = \mathbf{x}) = F\{\hat{h}^{[M]}(\upsilon|\mathbf{x})\} = F\left[\sum_{j=1}^{J} \{\mathbf{b}_j(\mathbf{x})^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon)^{\mathrm{T}}\} \boldsymbol{\gamma}_j^{[M]}\right].$$

Before we investigate the asymptotic properties of the resulting estimates, we shall discuss some details of this generic algorithm in what follows.

4.2. Model specification

The basis functions $\mathbf{b}_0$ and $\mathbf{b}_j$ determine the form of the fitted model, and their choice is problem specific. In the simplest situation, in which the conditional distribution of Y given only one numeric explanatory variable x_1 will be estimated, we could use the basis functions $\mathbf{b}_0(\upsilon) = (1, \upsilon)^{\mathrm{T}}$ and $\mathbf{b}_1(\mathbf{x}) = (1, x_1)^{\mathrm{T}}$. The corresponding base learner is then defined by the linear function $((1, x_1) \otimes (1, \upsilon))\boldsymbol{\gamma}_1 = (1, \upsilon, x_1, x_1\upsilon)\boldsymbol{\gamma}_1$. For each x_1, the transformation is linear in υ with intercept $\gamma_1 + \gamma_3 x_1$ and slope $\gamma_2 + \gamma_4 x_1$, i.e. not only the mean may depend on x_1 but also the variance. Restricting, for example, $\mathbf{b}_0(\upsilon)$ to be constant, i.e. $\mathbf{b}_0(\upsilon) \equiv 1$, allows the effects of explanatory variables to be restricted to the mean alone. Assuming that $\mathbf{b}_1(\mathbf{x}) \equiv 1$, however, yields a transformation function that is not affected by any explanatory variable. More flexible basis functions, e.g. B-spline basis functions, allow also for higher moments to depend on the explanatory variables. We illustrate appropriate choices of basis functions in Section 6.

4.3. Computational complexity

For the estimation of base learner parameters $\boldsymbol{\beta}_j$ in equation (6), it is not necessary to evaluate the Kronecker product $\otimes$ in expression (5) and to compute the $nN \times K_0 K_j$ design matrix for the jth base learner. The base learners that are used here are a special form of multi-dimensional smooth linear array models (Currie *et al.*, 2006), where efficient algorithms for computing Ridge estimates (6) exist. The number of multiplications required for fitting the jth base learner is approximately $c^6/(c^2/N - 1)$, instead of N^2c^4 for the simplest case with $c = K_0 = K_j$ and $N = n$ (see Table 2 in Currie *et al.* (2006)), and the memory required for storing the design matrices is of the order $NK_j + NK_0$, instead of NnK_jK_0. Note that only the gradient vector is of length Nn; all other objects can be stored in vectors or matrices growing with either N or n, and an explicit expansion of the observations $(Y_i \leqslant \upsilon_\iota, \mathbf{X}_i)$ for $i = 1, \ldots, N$ and $\iota = 1, \ldots, n$ is not necessary.

4.4. Choice of tuning parameters

The number of boosting iterations M is the most important tuning parameter determined by resampling, e.g. by k-fold cross-validation or bootstrapping. For the latter resampling scheme, the weights w_i in expression (4) are drawn from an N-dimensional multinomial distribution with constant probability parameters $p_i \equiv N^{-1}$, $i = 1, \ldots, N$. The out-of-bootstrap empirical risk with weights $w_i^{\mathrm{OOB}} = I(w_i = 0)$ is then used as a measure to assess the quality of the distributional forecasts for a varying number of boosting iterations M. The loss function that is used to fit the models is the same function as is used as a scoring rule to assess the quality of the probabilistic forecasts of the out-of-bootstrap observations.

The smoothing parameters λ_j, $j = 0, \ldots, J$, in the penalty matrices are not tuned but rather defined such that the jth base learner has low degrees of freedom. For our computations, we simplified the penalty term to $\mathbf{P}_{0j} = \lambda_j(\mathbf{P}_j \otimes \mathbf{1}_{K_0} + \mathbf{1}_{K_j} \otimes \mathbf{P}_0)$, i.e. one parameter controls the

smoothness in both directions. Following Hofner *et al.* (2011a), the parameters λ_j were defined such that each base learner has the same overall low degree of freedom. Note that the degree of freedom of the estimated partial transformation function adapts to the complexity that is inherent in the data via the number of boosting iterations M (Bühlmann and Yu, 2003). Different smoothness in the two directions can be imposed by choosing different basis functions for $\mathbf{b}_0$ and $\mathbf{b}_j$, e.g. a linear basis function for $\mathbf{b}_j$ and B-splines for $\mathbf{b}_0$. Other parameters, such as knots or degrees of basis functions or the number n of grid points that the integrated loss function l is approximated with are not considered as tuning parameters. The resulting estimates are quite insensitive to their different choices. Also, we do not consider the distribution function F or the loss function ρ as tuning parameter but assume that these are part of the model specification. Different versions of F and ρ lead to different negative gradients; these are given in Appendix A.

4.5. Monotonicity

The resulting estimate $\hat{h}^{[M]}(\upsilon|\mathbf{x})$ is not automatically monotone in its first argument. Monotonicity and smoothness in the Y-direction depend on each other, and too complex estimates tend to suffer from non-monotonicity. Empirically, on the basis of experiments that are reported in Sections 6 and 7, non-monotonicity is a problem in poorly fitting models, owing to either misspecification, overfitting or a low signal-to-noise ratio. From our point of view, inspecting the model for non-monotonicity is helpful for model diagnostics and can be dealt with by reducing model complexity. Alternatively, there are three possible modifications to the algorithm that can be implemented to enforce monotonicity:

(a) fit base learners under monotonicity constraints in equation (6), e.g. by using the iterative repenalization that was suggested by Eilers (2005),
(b) check monotonicity for each base learner and select the best among the monotone candidates only or
(c) select the base learner such that it is the best among all candidates that lead to monotone updates in $h^{[m]}$.

None of these approaches had to be used for our empirical studies, in which all resulting estimates were monotone for the appropriate number of boosting iterations M.

4.6. Model diagnostics and overfitting

Another convenient feature of transformation models is that, with the correct model h for absolute continuous random variables Y, the errors $E_i = h(Y_i|\mathbf{X}_i)$, $i = 1, \ldots, N$, are distributed according to F. Therefore, if the observed residuals $\hat{E}_i^{[M]} = \hat{h}^{[M]}(Y_i|\mathbf{X}_i)$ are unlikely to come from distribution F, e.g. assessed by using quantile–quantile plots or a Kolmogorov–Smirnov statistic, the model is likely to fit the data poorly. However, a good agreement between $\hat{E}_i^{[M]}$ and F does not necessarily mean that the explanatory variables describe the response well. A high correlation between the ranking of the residuals and the ranking of the responses $Y_1, \ldots, Y_N$ means that the estimated conditional distribution is very close to the unconditional empirical distribution of the responses. In this case, either the model may overfit or the response may be independent of the explanatory variables. The fitted model may also be used to draw novel responses for given explanatory variables by using the model-based bootstrap via $\tilde{Y}_i = \{\upsilon : Q(U_i) = \hat{h}^{[M]}(\upsilon|\mathbf{X}_i)\}$ for $i = 1, \ldots, N$, where $U_1, \ldots, U_N$ are independent and identically distributed uniform random variables. The stability of the model can now be investigated by refitting the model with observations $(\tilde{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, N$.

5. Consistency of boosted conditional transformation models

The boosting algorithm that is presented here is a variant of L_2WCBoost (Bühlmann and Yu, 2003) applied to dependent observations with more general base learners. In this section, we shall develop a consistency result for the squared error loss ρ_{sqe}. For simplicity, we consider the case in which the procedure is used with $F(h) = h$ as the identity function, i.e. the error term is uniformly distributed. Thus, we consider conditional transformation models of the form

$$\mathbb{P}(Y \leqslant \upsilon | \mathbf{X} = \mathbf{x}) = h(\upsilon|\mathbf{x}) = \sum_{j=1}^{J} h_j(\upsilon|x_j),$$

where the partial transformation function h_j is conditional on the jth explanatory variable in $\mathbf{x} = (x_1, \ldots, x_J) \in \chi$ and $\mathbb{E}_Y\{N^{-1}\Sigma_{i=1}^{N} h(Y|\mathbf{X}_i)\} = 0.5$. Our analysis is for the fixed design case with deterministic explanatory variables $\mathbf{X}_i$ or when conditioning on all $\mathbf{X}_i$s. A modification for the random-design case could be pursued along arguments that are similar to those for L_2-boosting as in Bühlmann (2006). As in Section 4, we use a basis expansion of $h(\upsilon|\mathbf{x})$:

$$h_{N,\gamma}(\upsilon|\mathbf{x}) = \sum_{j=1}^{J} \{\mathbf{b}_j(x_j)^{\mathrm{T}} \otimes \mathbf{b}_0(\upsilon)^{\mathrm{T}}\}\gamma_j = \sum_{j=1}^{J} \sum_{k_0=1}^{K_{0,N}} \sum_{k_1=1}^{K_{1,N}} \gamma_{j,k_0,k_1} b_{0,k_0}(\upsilon)\, b_{j,k_1}(x_j),$$

where for simplicity the number of basis functions $K_{1,N}$ is equal for all x_j.

Consider the (empirical) risk functions

$$R_{n,N}(h) = (nN)^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} \{I(Y_i \leqslant \upsilon_\imath) - h(\upsilon_\imath|\mathbf{X}_i)\}^2$$

and

$$R_{n,N,\mathbb{E}}(h) = (nN)^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} \mathbb{E}[\{I(Y_i \leqslant \upsilon_\imath) - h(\upsilon_\imath|\mathbf{X}_i)\}^2].$$

Denote the projected parameter by

$$\gamma_{0,N} = \arg\min_{\gamma} R_{n,N,\mathbb{E}}(h_{N,\gamma}). \tag{7}$$

We make the following assumptions.

Assumption 1. The coefficient vector $\gamma_{0,N}$ is sparse and satisfies

$$\|\gamma_{0,N}\|_1 = o\left[\sqrt{\left\{\frac{N}{\log(J_N K_{0,N} K_{1,N})}\right\}}\right], \qquad N \to \infty.$$

Thereby, the dimensionality $J = J_N$ can grow with N.

Assumption 2. The basis functions satisfy, for some $0 < C < \infty$,

$$\|b_{0,k_0}\|_\infty \leqslant C, \qquad \|b_{j,k_1}\|_\infty \leqslant C \qquad \forall j, k_0, k_1.$$

Assumption 3.

$$(nN)^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} h_{\gamma_{0,N}}(\upsilon_\imath|\mathbf{X}_i)^2 \leqslant D < \infty \qquad \forall n, N.$$

Assumption 1 is an l_1-norm sparsity assumption, assumption 2 is a mild restriction since we are modelling $I(Y \leqslant \upsilon)$ and assumption 3 requires that the signal strength does not diverge as $n, N \to \infty$.

Theorem 1. Assume assumptions 1–3. Then, for fixed n or for $n = n_N \to \infty$ ($N \to \infty$), and for $M = M_N \to \infty$ ($N \to \infty$), $M_N = o[\sqrt{\{N/\log(J_N K_{0,N} K_{1,N})\}}]$,

$$(nN)^{-1} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \{h_{\hat{\gamma}^{[M]}}(\upsilon_\iota|\mathbf{X}_i) - h_{\gamma_{0,N}}(\upsilon_\iota|\mathbf{X}_i)\}^2 = o_P(1), \qquad N \to \infty.$$

A proof is given in Appendix A.

Convergence of $h_{\gamma_{0,N}}(\upsilon|\mathbf{x})$ to the true function $h(\upsilon|\mathbf{x})$ involves approximation theory to achieve

$$(nN)^{-1} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \{h_{\gamma_{0,N}}(\upsilon_\iota|\mathbf{X}_i) - h(\upsilon_\iota|\mathbf{X}_i)\}^2 = o(1), \qquad n, N \to \infty. \tag{8}$$

We would want to estimate the function $h(\upsilon|\mathbf{x})$ well over the whole domain, e.g. $[a_\upsilon, b_\upsilon] \times \chi$. This may be too ambitious if $J = \dim(\chi) = J_N$ grows with N. Hence, we restrict ourselves to the setting where the number of active variables $J_{\text{act}} < \infty$ is fixed (from the active set S):

$$h(\upsilon|\mathbf{x}) = \sum_{j \in S} h_j(\upsilon|x_j), \qquad S \subseteq \{1, \ldots, J\} \text{ with } |S| = J_{\text{act}}.$$

For the approximation, we typically would need $K_{0,N}, K_{1,N} \to \infty$ ($N \to \infty$) for suitable basis functions and $n = n_N \to \infty$ ($N \to \infty$); furthermore, the grid $\upsilon_1 < \upsilon_2 < \ldots < \upsilon_n$ should become dense as $n = n_N \to \infty$, and also the values $\mathbf{X}_1^{\text{act}}, \ldots, \mathbf{X}_N^{\text{act}}$ should become dense in $\chi_S \subseteq \chi$ as $N \to \infty$ (here, $\mathbf{X}^{\text{act}} = \{X_j; j \in S\} \in \chi_S$). If $J = J_N$ grows, but the number of active variables in the model $J_{\text{act}} < \infty$ is fixed, then some uniform approximation $h_{\gamma_{0,N}}(\upsilon|\mathbf{x}) \to h(\upsilon|\mathbf{x})$ is possible under regularity conditions.

We provide a summary for a typical situation.

Corollary 1. Consider the setting as in theorem 1, with $J = J_N$ potentially growing but fixed dimensionality of the active variables $J_{\text{act}} < \infty$, $n = n_N \to \infty$ ($N \to \infty$), and the functions are sufficiently regular such that expression (8) holds. Then, for $M = M_N$ as in theorem 1,

$$(nN)^{-1} \sum_{i=1}^{N} \sum_{\iota=1}^{n} \{h_{\hat{\gamma}^{[M]}}(\upsilon_\iota|\mathbf{X}_i) - h(\upsilon_\iota|\mathbf{X}_i)\}^2 = o_P(1) \qquad n, N \to \infty.$$

This result states that the estimated $h_{\hat{\gamma}^{[M]}}$ are consistent for the true transformation h.

6. Applications

In this section, we present analyses with special emphasis on higher moments of the conditional distribution, which have received less attention in previous analyses of these problems. Further applications of conditional transformation models are given in Hothorn *et al.* (2012).

6.1. Childhood nutrition in India

Childhood undernutrition is one of the most urgent problems in developing and transition countries. To provide information not only on the nutritional status but also on health and population trends in general, demographic and health surveys conduct nationally representative surveys on fertility, family planning and maternal and child health, as well as child survival, human immunodeficiency virus–acquired immune deficiency syndrome, malaria and nutrition. Childhood nutrition is usually measured in terms of a Z-score that compares the nutritional status of children in the population of interest with the nutritional status in a reference

population. The nutritional status is expressed by anthropometric characteristics, i.e. height for age; in cases of chronic childhood undernutrition, the reduced growth rate in human development is termed stunted growth or stunting. The Z-score, which compares an anthropometric characteristic of child i with values from a reference population, is given as $Z_i = (\mathrm{AC}_i - m)/s$, where AC denotes the anthropometric characteristic of interest and m and s correspond to the median and (a robust estimate for the) standard deviation in the reference population (stratified with respect to age, gender and some other variables). We shall focus on stunting, i.e. insufficient AC equivalent to height for age, as a measure of chronic undernutrition in what follows and estimate the whole distribution of this Z-score measure for childhood nutrition in India. Our investigation is based on India's 1998–1999 Demographic and Health Survey (International Institute for Population Sciences and ORC Macro, 2000) on 24166 children visited during the survey in 412 of the 640 districts of India. The lower quantiles of this distribution can be used to assess the severity of childhood undernutrition, whereas the upper quantiles give us information about the nutritional status of children in families with above-average nutritional status.

The simplest conditional transformation model allowing for district-specific means and variances reads

$$\mathbb{P}(Z \leqslant v|\text{district} = k) = \Phi(\alpha_{0,k} + \alpha_k v), \qquad k = 1, \ldots, 412.$$

The base learner is defined by a linear basis $\mathbf{b}_0(v) = (1, v)^{\mathrm{T}}$ for the grid variable and a dummy encoding basis $\mathbf{b}_1(\text{district}) = (I(\text{district} = 1), \ldots, I(\text{district} = k))^{\mathrm{T}}$ for the 412 districts. The resulting 824-dimensional parameter vector γ_1 of the tensor product base learner then consists of separate intercept and slope parameters for each of the districts of India. Since we assume normality for the linear function $\alpha_{0,k} + \alpha_k Z \sim \mathcal{N}(0, 1)$, also the Z-score is assumed to be normal with both mean and variance depending on the district. We can relax the normal assumption on Z by allowing for more flexible transformations in the model

$$\mathbb{P}(Z \leqslant v|\text{district} = k) = \Phi\{h(v|\text{district} = k)\}, \qquad k = 1, \ldots, 412. \tag{9}$$

Now $\mathbf{b}_0(v)$ is a vector of B-spline basis functions evaluated at v for some reasonable choice of knots, whereas $\mathbf{b}_1$ remains as above. Hence, instead of assuming separate linear effects for the districts, we now assume separate non-parametric effects parameterized in terms of B-splines. To achieve smoothness of these non-parametric effects along the v-grid, we specify the penalty matrix $\mathbf{P}_0$ as $\mathbf{P}_0 = \mathbf{D}^{\mathrm{T}}\mathbf{D}$ with second-order difference matrix $\mathbf{D}$. It makes sense to induce spatial smoothness on the conditional distribution functions of neighbouring districts since we do not expect the distribution of the Z-score to change much from one district to its neighbouring districts. In fact, spatial smoothing is absolutely necessary in this example since otherwise we would estimate 412 separate distribution functions for the districts in India. To implement spatial smoothness of neighbouring districts, the penalty matrix $\mathbf{P}_1$ is chosen as an adjacency matrix, where the off-diagonal elements indicate whether two districts are neighbours (represented with a value of -1) or not (represented with a value of 0). The diagonal of the adjacency matrix contains the number of neighbours for the corresponding district. The estimated conditional transformation function $\hat{h}(Z|\text{district} = k)$ can be interpreted as a transformation of the Z-scores in district k to standard normality. Because the number of observations is large and the base learner is fitted with penalization, we stop the boosting algorithm when the reduction of the in-sample empirical risk is negligible.

From the estimated conditional distribution functions, we compute the τ-quantiles of the Z-score for each district via $\hat{Q}(\tau|\text{district} = k) = \inf[v : \Phi\{\hat{h}(v|\text{district} = k\} \geqslant \tau]$. The conditional 10% and 90% quantiles are depicted in a colour-coded map in Fig. 1. The spatially smooth

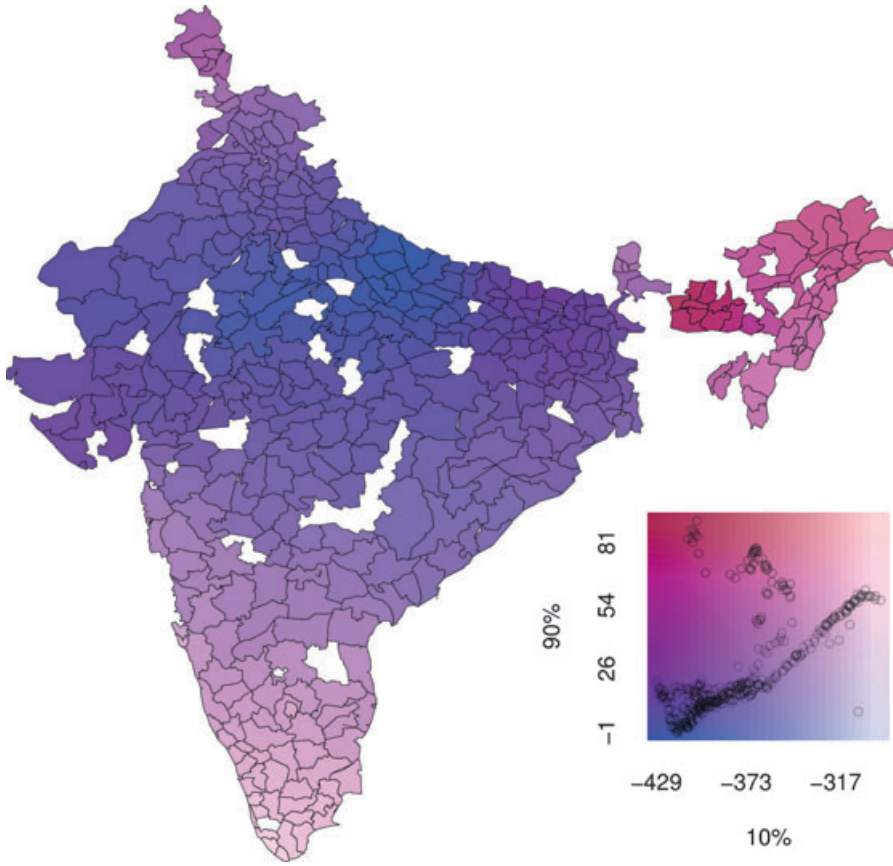

Fig. 1. Childhood nutrition in India—colour-coded map of the 10% and 90% conditional quantiles of the Z-score: each circle corresponds to one district with the respective colour in the map; blue values in the northern part of India correspond to small lower and upper quantiles; red values, especially in the eastern Meghalaya and Assam states, indicate small lower quantiles but at the same time large upper quantiles; in the southern part of India, the lower quantiles are largest with moderate upper quantiles; white parts indicate districts with no observations

estimated lower and upper conditional quantiles shown simultaneously allow differentiation between three groups of districts:

(a) districts with small lower and upper conditional quantiles (blue, especially in the Uttar Pradesh state), where the Z-score is stochastically smaller than that of the remaining parts of India and thus all children are less well fed;
(b) districts with more severe inequality, i.e. small lower but at the same time large upper quantiles (red, in the Meghalaya and Assam states);
(c) districts with relatively large lower and upper quantiles, which indicates a relatively good nutrition status of all children in the southern districts of India (violet, in Andhra Pradesh, Madhya Pradesh, Maharashtra, Tamil Nadu and Kerala).

6.2. Birth weight prediction

Recent advances in neonatal medicine have lowered the threshold of survival to a gestational age

of 23–24 weeks and to a birth weight of approximately 500 g. As neonatal risks of morbidity and mortality are highest in the lowest weight range, diagnostic assessment of the small fetus needs to be as precise as possible. Schild *et al.* (2008) focused on this high-risk group of small fetuses (1600 g and under) and proposed a formula for estimating birth weight BW based on ultrasound imaging performed within 7 days before delivery. In addition to predicting the expected birth weight given four standard two-dimensional ultrasound parameters (HC, head circumference; FE, femur length; BPD, biparietal diameter; AC, transverse diameter and circumference of the fetal abdomen) and three additional three-dimensional ultrasound parameters (UA, upper arm volume; FEM, thigh volume; ABDO, abdominal volume) $\mathbf{X} \in \mathbb{R}^7$, we aim at assessing the uncertainty of this prediction by 80% prediction intervals for birth weight (based on data of 150 predominantly Caucasian women collected in a prospective cohort study at the universities in Bonn and Erlangen, Germany; Schild *et al.* (2008)).

We begin with the linear model that was estimated by Schild *et al.* (2008),

$$\begin{aligned}\mathrm{BW}_{\mathrm{x}} = {} & 656.41 + 1.832\,\mathrm{ABDO} + 31.198\,\mathrm{HC} + 5.779\,\mathrm{FEM} + 73.521\,\mathrm{FL} + 8.301\,\mathrm{AC} \\ & - 449.886\,\mathrm{BPD} + 32.534\,\mathrm{BPD}^2 + 77.465\,\Phi^{-1}(U),\end{aligned}$$

and the classical prediction interval for a fetus with ultrasound parameters $\mathbf{x}$ is then the symmetric interval around the estimated conditional mean $\hat{\mathbb{E}}(\mathrm{BW}|\mathbf{X}=\mathbf{x})$, whose width is given by $2t_{150-8,0.9} \times 77.465\sqrt{[1 + \mathrm{var}\{\hat{\mathbb{E}}(\mathrm{BW}|\mathbf{X}=\mathbf{x})\}]}$.

The normality assumption can be relaxed by deriving the upper and lower conditional quantiles from two quantile regression models. Linear quantile regression (Koenker and Bassett, 1978) for the conditional 10%, 50% and 90% quantiles assumes that

$$\mathrm{BW}_{\mathbf{x}} = \alpha_{0,\tau} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_{\tau} + Q_{\tau}(U), \qquad \text{for } \tau = 0.1, \tau = 0.5 \text{ and } \tau = 0.9,$$

with $Q_{\tau}(\tau) = 0$. The corresponding prediction interval for a fetus with ultrasound parameters $\mathbf{x}$ is now $(\hat{\alpha}_{0,0.1} + \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{0.1}, \hat{\alpha}_{0,0.9} + \mathbf{x}^{\mathrm{T}}\hat{\boldsymbol{\alpha}}_{0.9})$. A more flexible description of the functional relationship between ultrasound parameters and quantiles is given by the additive quantile regression model (Koenker *et al.*, 1994)

$$\mathrm{BW}_{\mathbf{x}} = \alpha_{0,\tau} + \sum_{j=1}^{7} r_{j,\tau}(x_j) + Q_{\tau}(U), \qquad \text{for } \tau = 0.1, \tau = 0.5 \text{ and } \tau = 0.9.$$

Here, $r_{j,\tau}$ is a quantile-specific smooth function of the jth ultrasound parameter. Parameter tuning is difficult for these models; we therefore applied a boosting approach to additive quantile regression (Fenske *et al.* (2011), with early stopping via a 25-fold bootstrap). Prediction intervals can now be derived by $\{\hat{\alpha}_{0,0.1} + \Sigma_{j=1}^{7}\hat{r}_{j,0.1}(x_j), \hat{\alpha}_{0,0.9} + \Sigma_{j=1}^{7}\hat{r}_{j,0.9}(x_j)\}$. Note that, for either quantile regression model, the prediction interval is based on two separate models: one for $\tau = 0.1$ and one for $\tau = 0.9$.

Finally, we derive prediction intervals from the conditional transformation model

$$\mathbb{P}(\mathrm{BW} \leqslant \upsilon|\mathbf{X}=\mathbf{x}) = \Phi\{h(\upsilon|\mathbf{x})\} = \Phi\left\{\sum_{j=1}^{7} h_j(\upsilon|x_j)\right\}$$

where, under the assumption of additivity of the transformation function h, each ultrasound parameter may influence the moments of the conditional birth weight distribution. The jth base learner is the tensor product of B-spline basis functions $\mathbf{b}_0(\upsilon)$ for birth weight and $\mathbf{b}_j(x_j)$ are B-spline basis functions for the jth ultrasound parameter. The penalty matrices $\mathbf{P}_0$ and $\mathbf{P}_j$ penalize second-order differences, and thus all estimates $\hat{h}_j$ will be smooth bivariate tensor

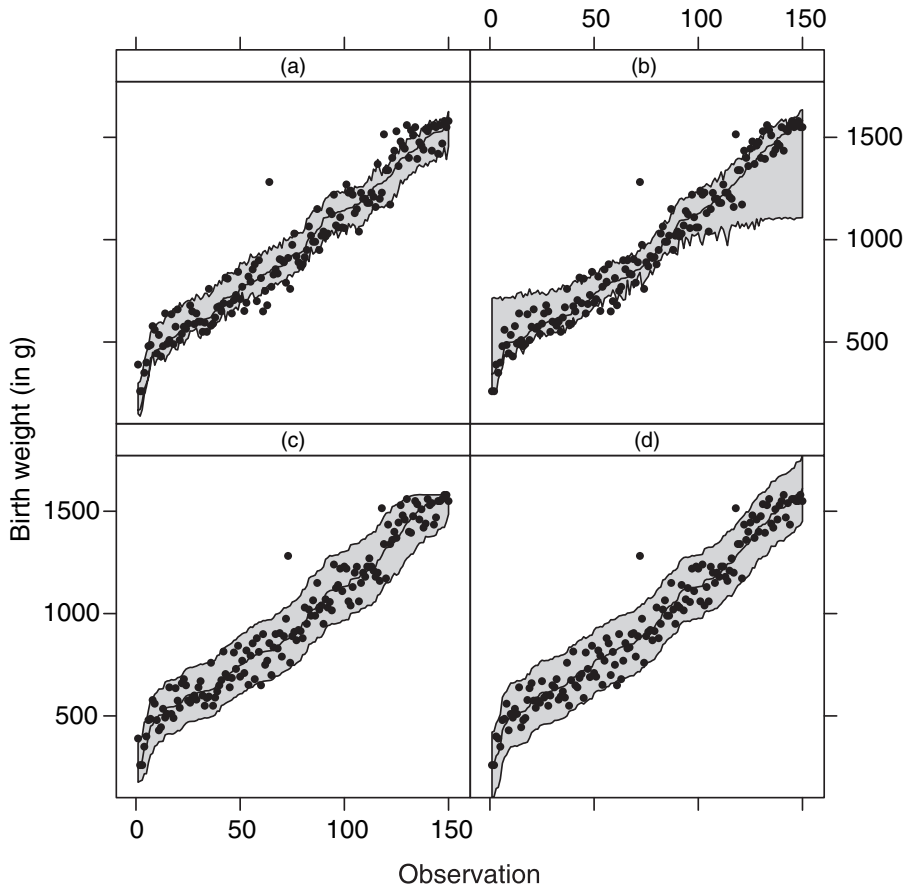

Fig. 2. Birth weight prediction—observed birth weights of 150 small fetuses (•) ordered with respect to the estimated mean or median expected birth weight (———): ▭, fetus-specific 80% prediction intervals for (a) the linear quantile regression model, (b) the additive quantile regression model, (c) the conditional transformation model and (d) the linear model

product splines of birth weight and the respective ultrasound parameter, with both dimensions being subject to smoothing. The number of boosting iterations was determined by a 25-fold bootstrap. From the estimated conditional distribution functions, we compute the τ-quantiles of the birth weight via $\hat{Q}(\tau|\mathbf{X}=\mathbf{x}) = \inf[\upsilon : \Phi\{h(\upsilon|\mathbf{x})\} \geqslant \tau]$ and derive the prediction interval as $(\hat{Q}(0.1|\mathbf{X}=\mathbf{x}), \hat{Q}(0.9|\mathbf{X}=\mathbf{x}))$.

The observed birth weights ordered with respect to the predicted mean (linear model) or median (quantile regression and conditional transformation model) are depicted in Fig. 2. In addition, the respective 80% prediction intervals are visualized by grey areas. It should be noted that, for all models, the prediction intervals are only interpretable for future observations; however, poor coverage for the learning sample also indicates poor coverage for future cases. The prediction intervals that were obtained from linear quantile regression indicate that the model is confident about its predictions over the whole range of birth weights. This is also so for the additive quantile regression models for birth weights of approximately 1000 g, but the uncertainty increases for very small and larger fetuses. The intervals that were obtained from the linear model and the conditional transformation model are similar. For birth weights

between 500 and 1400 g, the prediction intervals of the conditional transformation model are symmetric around the median. This might be an indication that the normality assumption by the linear model is not completely unrealistic. The smaller interval widths that can be seen for the linear model are most likely to be due to the variance estimate in this case ignoring the model choice process that was performed before the final model fit by Schild *et al.* (2008). The conditional transformation model takes this variability into account. The results may also be an indication that the assumption of additivity of the transformation function rather than of the regression function (for quantile regression models) might be more appropriate for modelling birth weights.

7. Empirical evaluation

We shall compare the empirical performance of conditional transformation models fitted by means of the proposed boosting algorithm to two competitors. Conditional transformation models are semiparametric models in the sense that we assume a certain distribution for the transformed responses and additivity of the model terms on the scale of the corresponding quantile function. Therefore, it is natural to compare these estimated conditional distribution functions with a fully parametric approach and a non-parametric estimation technique.

For simplicity, we study a model in which two explanatory variables influence both the conditional expectation and the conditional variance of a normally distributed response Y. The error term $\Phi^{-1}(U)$ is standard normal, and, to obtain normal responses, we restrict the possible transformations to linear functions:

$$\Phi^{-1}(U) = h(Y_{\mathbf{x}}|\mathbf{x}) = \sum_j h_j(Y_{\mathbf{x}}|\mathbf{x}) = \sum_j b_j(\mathbf{x}) Y_{\mathbf{x}} - a_j(\mathbf{x}) = Y_{\mathbf{x}} \sum_j b_j(\mathbf{x}) - \sum_j a_j(\mathbf{x})$$

$$\Leftrightarrow Y_{\mathbf{x}} = \frac{\Phi^{-1}(U) + \sum_j a_j(\mathbf{x})}{\sum_j b_j(\mathbf{x})} \sim \mathcal{N}\left[\frac{\sum_j a_j(\mathbf{x})}{\sum_j b_j(\mathbf{x})}, \left\{ \sum_j b_j(\mathbf{x}) \right\}^{-2} \right].$$

Although the partial transformation functions are linear in $Y_{\mathbf{x}}$, the expectation and variance depend on the explanatory variables in a non-linear way. The choices $X_1 \sim \mathcal{U}[0, 1]$, $X_2 \sim \mathcal{U}[-2, 2]$, $a_1(\mathbf{x}) = 0$, $a_2(\mathbf{x}) = x_2$, $b_1(\mathbf{x}) = x_1$ and $b_2(\mathbf{x}) = 0.5$ lead to the heteroscedastic varying-coefficient model

$$Y_{\mathbf{x}} = \frac{1}{x_1 + 0.5} x_2 + \frac{1}{x_1 + 0.5} \Phi^{-1}(U), \tag{10}$$

where the variance of $Y_{\mathbf{x}}$ ranges between 0.44 and 4 depending on X_1. This model can be fitted in the GAMLSS framework under the assumptions that the expectation of the normal response depends on a smoothly varying regression coefficient $(X_1 + 0.5)^{-1}$ for X_2 and that the variance is a smooth function of X_1. This model is therefore fully parametric. As a non-parametric counterpart, we use a kernel estimator for estimating the conditional distribution function of $Y_{\mathbf{x}}$ as a function of the two explanatory variables.

The conditional transformation model

$$\mathbb{P}(Y \leqslant v | X_1 = x_1, X_2 = x_2) = \Phi\{h(v|x_1, x_2)\} = \Phi\{h_1(v|x_1) + h_2(v|x_2)\}$$

is a semiparametric compromise between these two extremes. The error distribution is assumed to be standard normal and additivity of the transformation function h is also part of the model specification. The base learners are tensor products of B-spline basis functions $\mathbf{b}_0(v)$ for Y and

B-spline basis functions for X_1 and X_2. The penalty matrices $\mathbf{P}_0$, $\mathbf{P}_1$ and $\mathbf{P}_2$ penalize second-order differences, and thus $\hat{h}_j$ will be smooth bivariate tensor product splines of the response and explanatory variables X_1 and X_2. Smoothing takes place in both dimensions.

For all three approaches, we obtain estimates of $\mathbb{P}(Y \leqslant \upsilon | X_1 = x_1, X_2 = x_2)$ over a grid on x_1 and x_2 and compute the mean absolute deviation MAD of the true and estimated probabilities,

$$\text{MAD}(x_1, x_2) = \frac{1}{n} \sum_{\upsilon} |\mathbb{P}(Y \leqslant \upsilon | X_1 = x_1, X_2 = x_2) - \hat{\mathbb{P}}(Y \leqslant \upsilon | X_1 = x_1, X_2 = x_2)|,$$

for each pair of x_1 and x_2. Then, the minimum, the median and the maximum of the MAD-values over this grid are computed as summary statistics. This procedure was repeated for 100 random samples of size $N = 200$ drawn from model (10). Cross-validation was used to determine the bandwidths for the kernel-based methods; for details see Hayfield and Racine (2008). The boosting-based estimation of GAMLSSs (Mayr *et al.*, 2012a) turned out to be more stable than the reference implementation (package `gamlss`; Stasinopoulos *et al.* (2011)), and we therefore fitted the GAMLSSs by the dedicated boosting algorithm. For GAMLSSs and conditional transformation models fitted by boosting, the number of boosting iterations was determined via sample splitting. To investigate the stability of the three procedures under non-informative explanatory variables, we added $p = 1, \ldots, 5$ uniformly distributed variables without association to the response to the data and included them as potential explanatory variables in the three models. The case $p = 0$ corresponds to model (10).

Fig. 3 shows the empirical distributions of the minimum, median and maximum MAD for the three competitors. For $p = 0$, the GAMLSS and conditional transformation models perform on par with respect to the median MAD, although the GAMLSS shows a somewhat larger variability. The median MAD is slightly smaller than 0.02 for both procedures, which indicates that the true conditional distribution function can be fitted precisely. The maximal MAD is smallest for conditional transformation models and can be quite large for the GAMLSS. In contrast, for some configurations of the explanatory variables, the GAMLSS seems to offer better estimates with respect to the minimal MAD. The kernel estimator leads to the largest median MAD-values but seems more robust than the GAMLSS with respect to the maximal MAD. These results are remarkably robust in the presence of up to five non-informative explanatory variables, although of course the MAD increases with p.

The general theme that the GAMLSS on average performs as well as conditional transformation models in the special case of model (10) but is associated with a larger variability might be explained by the independent estimation of the functions for the expectation and variance, i.e. the GAMLSS does not 'know' that the varying-coefficient term and the variance term are actually the same. The inferior performance of the kernel estimator might be explained by the technical difficulties that are associated with bandwidth choice. The tuning parameters for the two boosting approaches are easier to choose. Our general impression is that the kernel-estimated conditional distribution functions are more erratic than the smooth functions that are obtained with boosting for conditional transformation models (the analysis of the simulation data is not shown here).

Since conditional transformation models are also an alternative to quantile regression models, it would be interesting to compare the two approaches. At this point, it is important to recall that the two models assume additivity of the effects of X_1 and X_2, but on different scales as explained in Section 2. Consequently, the heteroscedastic varying-coefficient model (10) cannot be fitted in a straightforward way by using standard linear or additive quantile regression. However, the estimation problem can be slightly reformulated by describing the τ-quantile of $Y_{\mathbf{x}}$ as the sum of a varying-coefficient term $r_1(x_1)x_2$ and a smooth function $r_2(x_1)$. This model, implemented

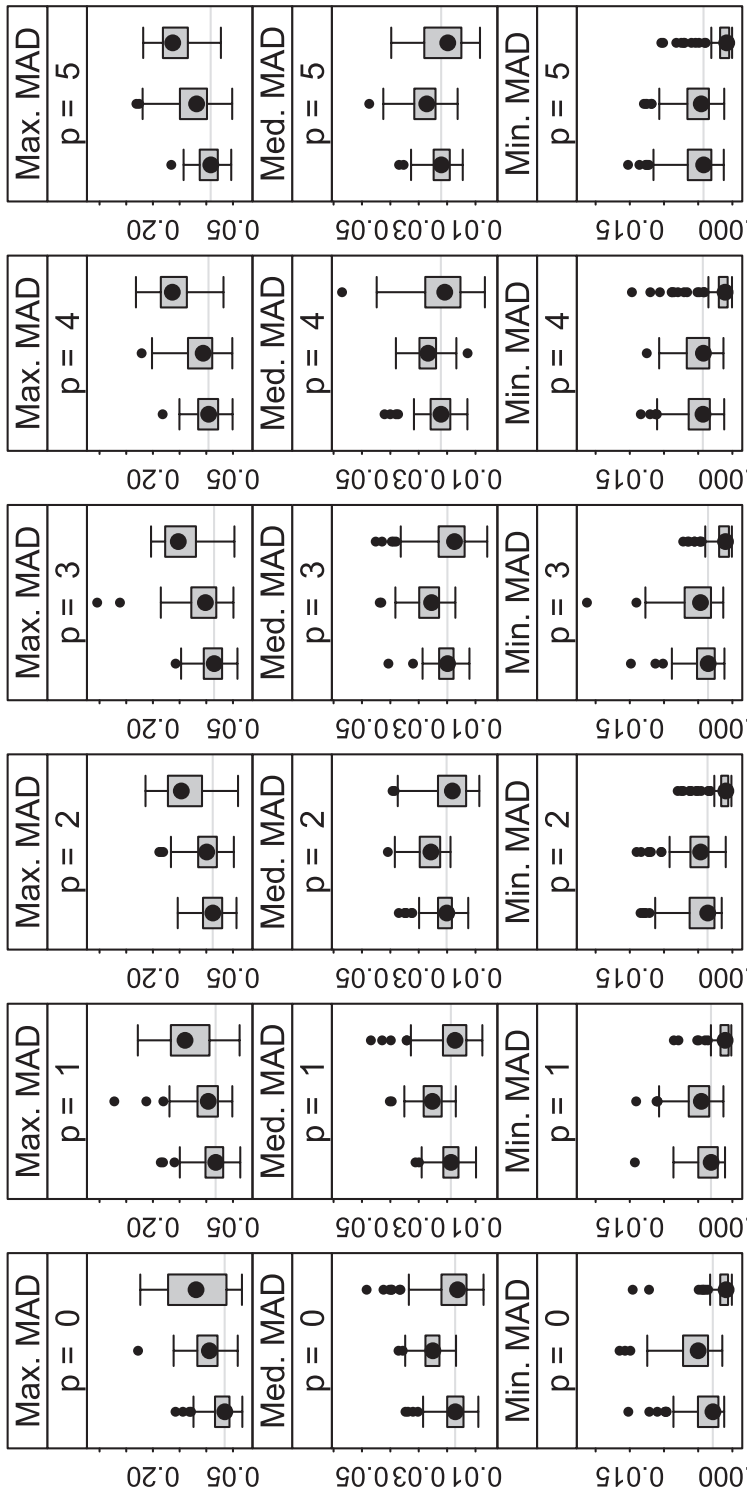

Fig. 3. Empirical evaluation—minimum, median and maximum of the mean absolute deviation MAD between the true and estimated probabilities for conditional transformation models (CTM), non-parametric kernel distribution function estimation and GAMLSSs for 100 random samples: values on the ordinate can be interpreted as absolute differences of probabilities; ——, median of the conditional transformation models

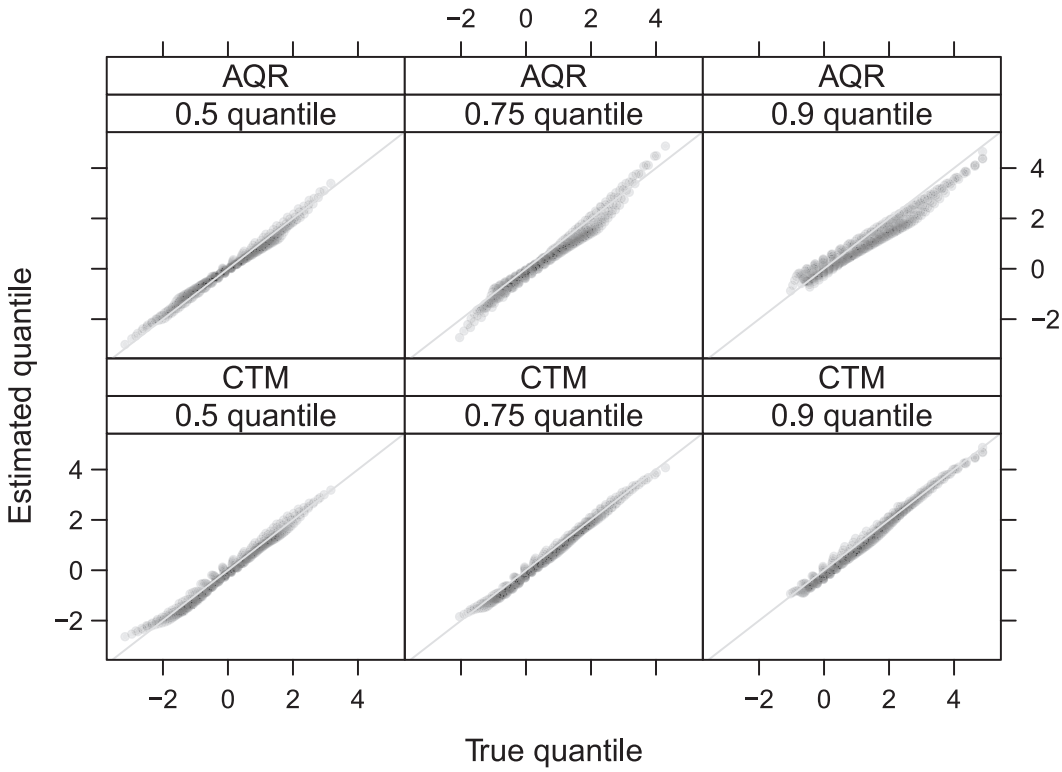

Fig. 4. Comparison of additive quantile regression (AQR) and conditional transformation models (CTM): scatter plots of true *versus* estimated quantiles obtained from one conditional transformation model and from three additive quantile regression models fitted to 200 observations drawn from the heteroscedastic varying-coefficient model (10)

by using the boosting approach to additive quantile regression with varying coefficients that was introduced by Fenske *et al*. (2011), allows the estimation of conditional τ-quantiles.

We fitted three such quantile regression models (for $\tau = 0.5, 0.75, 0.9$) to a sample of size $N = 200$ from model (10) and determined the optimal number of boosting iterations by the out-of-sample empirical risk of the check function. To give an impression, we compare these estimated τ-quantiles with the corresponding conditional quantiles obtained by inverting the estimated conditional distribution function from a conditional transformation model. Fig. 4 displays scatter plots of the true conditional quantiles over a grid of x_1- and x_2-values with the corresponding estimated quantiles derived from one conditional transformation model and the three additive quantile regression models for $\tau = 0.5, 0.75, 0.9$; the latter models include the varying-coefficient term. It seems that, in this example, both approaches recover the true quantiles equally well.

8. Discussion

In *Quantile Regression*, Koenker (2005) put transformation models in the 'twilight zone of quantile regression' and suggested that estimating conditional distribution functions by means of transformation models might be an alternative to the direct estimation of conditional quantile functions. We undertook the 'worthwhile exercise' (Koenker (2005), section 8.1.1) and devel-

oped a semiparametric framework for the estimation of conditional distribution functions by conditional transformation models that allows higher moments of the conditional distribution to depend on the explanatory variables.

Because the empirical risk function (4) is equivalent to well-established risk functions for binary data, many potentially interesting algorithms can be used to fit conditional transformation models, although dependent observations must be dealt with. We chose a componentwise boosting approach mainly because of its 'divide-and-conquer' strategy, which allows a very efficient fitting of base learners that depend on the response and on one or more explanatory variables at the same time via linear array models. Although boosting became popular owing to its success in fitting simple models under challenging circumstances—especially linear or additive models for high dimensional explanatory variables—the attractiveness of this class of algorithms for fitting challenging models in simple circumstances has been only rarely recognized. Exceptions are Ridgeway (2002) and Sexton and Laake (2012), who studied boosting algorithms for fitting density functions. Lu and Li (2008), Schmid and Hothorn (2008) and Schmid *et al*. (2011) proposed boosting algorithms for transformation models that treat the transformation function h_Y as a nuisance parameter. In the same model framework, Tutz and Groll (2012) proposed a likelihood boosting approach for fitting cumulative and sequential models for ordinal responses.

Boosting algorithms for estimating conditional quantiles by minimizing the check function have been introduced by Kriegler and Berk (2010), Fenske *et al*. (2011) and Zheng (2012). The computation of prediction intervals based on pairs of such models is quite straightforward (Mayr *et al*., 2012b). Our approach to the estimation of the conditional distribution function has the advantage that one model fits the whole distribution, which can then be used to derive arbitrary functionals. The quantile score representation of the continuous ranked probability score (see Gneiting and Ranjan (2011)) might be a basis to develop a boosting technique which is similar to that described in this paper for the estimation of full conditional quantile functions. The main difference between transformation and quantile regression models that we must keep in mind is that additivity is assumed on two different scales. From a practical point of view, diagnostic tools to assess which of these scales is more appropriate for assuming an additive model would be very important.

The applications that were presented in Section 6 showed that conditional transformation models are generic, and we can, by choosing appropriate base learners, fit models that are specific to the problem at hand. An empirical evaluation showed that the estimated conditional distribution functions are on average as good as the estimates that are obtained from a parametric approach (the GAMLSS) that relies on more assumptions. In comparison with non-parametric kernel distribution estimators, conditional transformation models are more adaptable, for example, to spatial or temporal data. The performance of the semiparametric models compared with that of the non-parametric competitor was considerably better at the small price of the assumption of additivity of the transformation function.

It will be interesting to study conditional transformation models further with respect to the following extensions. Discrete distributions can be handled by basis functions $\mathbf{b}_0$ offering one parameter for each element of the support of Y, similarly to a proportional odds model (see Hothorn *et al*. (2012) for an application). Instead of making assumptions about the quantile function Q representing the error distribution, it would be possible to fit the corresponding distribution function by techniques introduced for single-index models (Tutz and Petry, 2012). Accelerated failure time models fitted by boosting of an inverse probability of censoring weighted risk have been described by Hothorn *et al.* (2006), and future research awaits the investigation of the performance of conditional transformation models under censoring.

9. Computational details

Conditional transformation models were fitted by using an implementation of componentwise boosting in package `mboost` (version 2.1-2; Hothorn *et al.* (2011)). Package `gamboostLSS` (version 1.0-3; Hofner *et al.* (2011b)) was used to fit GAMLSSs and kernel distribution estimation was performed by using package `np` (version 0.40-13; Hayfield and Racine (2011)). Linear quantile regression was computed by using package `quantreg` (version 4.79; Koenker (2011)). All computations were performed by using R version 2.13.2 (R Development Core Team, 2011).

Throughout Section 6, we used the loss function that was defined by ρ_{bin} and modelled non-linear functions by cubic B-spline bases with 20 equidistant knots. For further computational details we refer the reader to the R code that implements the analyses that were presented in Sections 6 and 7, which is available in an experimental R package `ctm` at `http://R-forge.R-project.org/projects/ctm`. The results that are presented in this paper can be reproduced by using this package, except for the birth weight data, which are not publicly available.

Acknowledgements

We thank the Joint Editor, the Associate Editor and a referee for their careful review of an initial version of this paper. Andreas Mayr implemented the `Normal()` family for `gamboostLSS` by our request, Achim Zeileis helped to find the right colours for Fig. 1 and Ronald Schild provided us with the birth weight data. We are indebted to Paul Eilers, Tilman Gneiting and Roger Koenker for their comments on a draft version and we thank Karen A. Brune for improving the language. Financial support by the Deutsche Forschungsgemeinschaft (grant HO 3242/4-1) is gratefully acknowledged.

Appendix A: Proofs

A.1. Proof of lemma 1

A.1.1. Convexity

If the loss function ρ is convex in its second argument, so is the loss function l,

$$l\{(Y,\mathbf{X}),\alpha h+(1-\alpha)g\}\leqslant\alpha\, l\{(Y,\mathbf{X}),h\}+(1-\alpha)\, l\{(Y,\mathbf{X}),g\},$$

with $g(\cdot|\mathbf{x}):\mathbb{R}\to\mathbb{R}$ being a monotone increasing transformation function and $\alpha\in[0,1]$, because of the convexity of ρ and the monotonicity and linearity of the Lebesgue integral.

A.1.2. Population minimizers

Let f denote the density of F. With iterated expectation we have

$$\begin{aligned}\mathbb{E}_{Y,\mathbf{X}}[l\{(Y,\mathbf{X}),h\}] &= \int\int\int \rho\{(y\leqslant v,\mathbf{x}),h(v|\mathbf{x})\}\,\mathrm{d}\mu(v)\,\mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y)\,\mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x})\\ &= \int\int\underbrace{\int \rho\{(y\leqslant v,\mathbf{x}),h(v|\mathbf{x})\}\,\mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y)}_{=:A_{v,\mathbf{x}}\{h(v|\mathbf{x})\}}\,\mathrm{d}\mu(v)\,\mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x})\end{aligned}$$

and the risk is minimal when $A_{v,\mathbf{x}}\{h(v|\mathbf{x})\}$ is minimal for the scalar $h(v|\mathbf{x})$ for all v and $\mathbf{x}$, i.e. when

$$0\stackrel{!}{=}\frac{\partial A_{v,\mathbf{x}}\{h(v|\mathbf{x})\}}{\partial\{h(v|\mathbf{x})\}}$$

$$
\begin{aligned}
&= \int \frac{\partial}{\partial h(\upsilon|\mathbf{x})} \rho\{(y \leqslant \upsilon, \mathbf{x}), h(\upsilon|\mathbf{x})\} \, \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&\overset{\rho=\rho_{\text{sqe}}}{=} \int [I(y \leqslant \upsilon) - F\{h(\upsilon|\mathbf{x})\} \, f\{h(\upsilon|\mathbf{x})\}] \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&= f\{h(\upsilon|\mathbf{x})\} \left[\int I(y \leqslant \upsilon) \, \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) - F\{h(\upsilon|\mathbf{x})\} \right] \\
&= f\{h(\upsilon|\mathbf{x})\} [\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) - F\{h(\upsilon|\mathbf{x})\}]
\end{aligned}
$$

which for $f\{h(\upsilon|\mathbf{x})\} > 0$ is 0 for $h(\upsilon|\mathbf{x}) = F^{-1}\{\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})\}$. Similar, for $\rho = \rho_{\text{bin}}$ the term

$$
\begin{aligned}
0 &\overset{!}{=} \frac{\partial A_{\upsilon, \mathbf{x}}\{h(\upsilon|\mathbf{x})\}}{\partial h(\upsilon|\mathbf{x})} \\
&= \int -\left[\frac{I(y \leqslant \upsilon)}{F\{h(\upsilon|\mathbf{x})\}} f\{h(\upsilon|\mathbf{x})\} - \frac{1 - I(y \leqslant \upsilon)}{1 - F\{h(\upsilon|\mathbf{x})\}} f\{h(\upsilon|\mathbf{x})\} \right] \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&= f\{h(\upsilon|\mathbf{x})\} \left[\frac{\int 1 - I(y \leqslant \upsilon) \, \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y)}{1 - F\{h(\upsilon|\mathbf{x})\}} - \frac{\int I(y \leqslant \upsilon) \, \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y)}{F\{h(\upsilon|\mathbf{x})\}} \right] \\
&= f\{h(\upsilon|\mathbf{x})\} \left[\frac{1 - \mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})}{1 - F\{h(\upsilon|\mathbf{x})\}} - \frac{\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})}{F\{h(\upsilon|\mathbf{x})\}} \right]
\end{aligned}
$$

is 0 for $h(\upsilon|\mathbf{x}) = F^{-1}\{\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})\}$ when $f\{h(\upsilon|\mathbf{x})\} > 0$.

For the absolute error, note that

$$
\rho_{\text{abe}}\{(Y \leqslant \upsilon, \mathbf{X}), h(\upsilon|\mathbf{X})\} = I(Y \leqslant \upsilon)[1 - F\{h(\upsilon|\mathbf{X})\}] + \{1 - I(Y \leqslant \upsilon)\} F\{h(\upsilon|\mathbf{X})\}
$$

and thus

$$
\begin{aligned}
A_{\upsilon, \mathbf{x}}\{h(\upsilon|\mathbf{x})\} &= \int \rho_{\text{abe}}\{(y \leqslant \upsilon, \mathbf{x}), h(\upsilon|\mathbf{x})\} \, \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&= \int I(Y \leqslant \upsilon)[1 - F\{h(\upsilon|\mathbf{X})\}] + \{1 - I(Y \leqslant \upsilon)\} F\{h(\upsilon|\mathbf{X})\} \, \mathrm{d}\mathbb{P}_{Y|\mathbf{X}=\mathbf{x}}(y) \\
&= [1 - F\{h(\upsilon|\mathbf{X})\}] \mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) + F\{h(\upsilon|\mathbf{X})\}\{1 - \mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})\}.
\end{aligned}
$$

This expression attains its minimal value of $\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})$ for $\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) \leqslant 0.5$ when $F\{h(\upsilon|\mathbf{X})\} = 0$. For $\mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x}) > 0.5$, the minimum $1 - \mathbb{P}(Y \leqslant \upsilon|\mathbf{X}=\mathbf{x})$ is attained when $F\{h(\upsilon|\mathbf{X})\} = 1$. Thus, absolute error will lead to too extreme estimated values of h and corresponding conditional distribution functions.

A.2. Proof of theorem 1

We use a modified argument of a proof that was presented in section 12.8.2. in Bühlmann and van de Geer (2011). Formally, we can write

$$
\begin{aligned}
I(Y_i \leqslant \upsilon_\iota) &= h_{\gamma_{0,N}}(\upsilon_\iota|\mathbf{X}_i) + \varepsilon_{i\iota}, \\
\varepsilon_{i\iota} = I(Y_i \leqslant \upsilon_\iota) - h_{\gamma_{0,N}}(\upsilon_\iota|\mathbf{X}_i) &\qquad i = 1, \ldots, N, \quad \iota = 1, \ldots, n.
\end{aligned}
$$

The errors $\varepsilon_{i\iota}$ have reasonable properties, as discussed in equation (11) below.

There are two issues that need to be addressed. First, we define the inner products of functions h and g:

$$
(h, g)_{n,N,\mathbb{E}} = n^{-1} \sum_{\iota=1}^{n} \mathbb{E}\{h(\upsilon_\iota|\mathbf{X}) \, g(\upsilon_\iota|\mathbf{X})\}
$$

and

$$
(h, g)_{n,N} = n^{-1} N^{-1} \sum_{i=1}^{N} \sum_{\iota=1}^{n} h(\upsilon_\iota|\mathbf{X}_i) \, g(\upsilon_\iota|\mathbf{X}_i).
$$

The proof in Bühlmann and van de Geer (2011) can then be used with the scalar product $(h, g)_{n,N}$.

Secondly, for controlling the probabilistic part of the proof, we need to show that the analogue of formula (12.26) in Bühlmann and van de Geer (2011) holds. This translates to deriving a bound for

$$\max_{j,k_0,k_1} (b_{0,k_0} b_{j,k_1}, \varepsilon)_{n,N} = \max_{j,k_0,k_1} (nN)^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} b_{0,k_0}(\upsilon_\imath)\, b_{j,k_1}(X_j) \varepsilon_{i\imath}.$$

Because $h_{\gamma_{0,N}}$ is the projection of $I(Y \leqslant \upsilon_\imath)$, $\imath = 1, \ldots, n$, onto the basis functions $b_{0,k_0}(\upsilon_\imath)\, b_{j,k_1}(X_j)$, $\imath = 1, \ldots, n$, with respect to the $\|\cdot\|_{n,N,\mathbb{E}}$-norm (see expression (7)), and, owing to the definition of $\varepsilon_{i\imath}$, we have

$$n^{-1} \sum_{\imath=1}^{n} \mathbb{E}\{\varepsilon_{i\imath}\, b_{0,k_0}(\upsilon_\imath) b_{j,k_1}(X_j)\} = 0 \qquad \forall j, k_0, k_1. \tag{11}$$

Therefore,

$$(nN)^{-1} \sum_{i=1}^{N} \sum_{\imath=1}^{n} b_{0,k_0}(\upsilon_\imath)\, b_{j,k_1}(X_j) \varepsilon_{i\imath} = N^{-1} \sum_{i=1}^{N} Z_i(j, k_0, k_1),$$
$$\mathbb{E}\{Z_i(j, k_0, k_1)\} = 0.$$

Furthermore, owing to the boundedness assumption in assumption 1, $\|Z_i(j, k_0, k_1)\|_\infty \leqslant C_1$ for some constant $C_1 < \infty$, $\forall i, j, k_0, k_1$. Applying Hoeffding's inequality, for independent (but not necessarily identically distributed) random variables (van de Geer (2000), lemma 3.5) and using the union bound, we obtain

$$\max_{j,k_0,k_1} (b_{0,k_0} b_{j,k_1}, \varepsilon)_{n,N} = O_P[\sqrt{\{\log(J_N K_{0,N} K_{1,N})/N\}}].$$

This, together with the proof from section 12.8.2 in Bühlmann and van de Geer (2011), completes the proof of theorem 1.

A.3. Gradients

We present the gradients for different loss functions ρ and arbitrary absolute continuous distribution functions F with density function f:

$$U_{i\imath} \overset{\rho=\rho_{\text{bin}}}{=} \left\{ \frac{I(Y_i \leqslant \upsilon_\imath)}{F(\hat{h}_{i\imath}^{[m]})} - \frac{1 - I(Y_i \leqslant \upsilon_\imath)}{1 + F(\hat{h}_{i\imath}^{[m]})} \right\} f(\hat{h}_{i\imath}^{[m]}),$$
$$U_{i\imath} \overset{\rho=\rho_{\text{sqe}}}{=} \{I(Y_i \leqslant \upsilon_\imath) - F(\hat{h}_{i\imath}^{[m]})\}\, f(\hat{h}_{i\imath}^{[m]}),$$
$$U_{i\imath} \overset{\rho=\rho_{\text{abe}}}{=} [I(Y_i \leqslant \upsilon_\imath) - \{1 - I(Y_i \leqslant \upsilon_\imath)\}]\, f(\hat{h}_{i\imath}^{[m]}).$$

References

Bühlmann, P. (2006) Boosting for high-dimensional linear models. *Ann. Statist.*, **34**, 559–583.

Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications*. Berlin: Springer.

Bühlmann, P. and Hothorn, T. (2007) Boosting algorithms: regularization, prediction and model fitting (with discussion). *Statist. Sci.*, **22**, 477–505.

Bühlmann, P. and Yu, B. (2003) Boosting with the L_2 loss: regression and classification. *J. Am. Statist. Ass.*, **98**, 324–339.

Chen, K. and Müller, H. G. (2012) Conditional quantile analysis when covariates are functions, with application to growth data. *J. R. Statist. Soc.* B, **74**, 67–89.

Chen, K. and Tong, X. (2010) Varying coefficient transformation models with censored data. *Biometrika*, **97**, 969–976.

Cheng, G. and Wang, X. (2011) Semiparametric additive transformation model under current status data. *Electron. J. Statist.*, **5**, 1735–1764.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika*, **82**, 835–845.

Currie, I. D., Durban, M. and Eilers, P. H. C. (2006) Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc.* B, **68**, 259–280.

Dette, H. and Volgushev, S. (2008) Non-crossing non-parametric estimates of quantile curves. *J. R. Statist. Soc.* B, **70**, 609–627.

Doksum, K. A. and Gasko, M. (1990) On a correspondence between models in binary regression analysis and in survival analysis. *Int. Statist. Rev.*, **58**, 243–252.
Eilers, P. H. C. (2005) Unimodal smoothing. *J. Chemometr.*, **19**, 317–328.
Fenske, N., Kneib, T. and Hothorn, T. (2011) Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J. Am. Statist. Ass.*, **106**, 494–510.
Friedman, H. H., Friedman, L. W. and Amoo, T. (2002) Using humor in the introductory statistics course. *J. Statist. Educ.*, **10**, no. 3.
van de Geer, S. (2000) *Empirical Processes in M-estimation*. Cambridge: Cambridge University Press.
Gilchrist, W. (2008) Regression revisited. *Int. Statist. Rev.*, **76**, 401–418.
Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.
Gneiting, T., Raftery, A. E., Westveld III, A. H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mnthly Weath. Rev.*, **133**, 1098–1118.
Gneiting, T. and Ranjan, R. (2011) Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Statist.*, **29**, 411–422.
Hall, P. and Müller, H. G. (2003) Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *J. Am. Statist. Ass.*, **98**, 598–608.
Hall, P., Wolff, R. C. L. and Yao, Q. (1999) Methods for estimating a conditional distribution function. *J. Am. Statist. Ass.*, **94**, 154–163.
Hayfield, T. and Racine, J. S. (2008) Nonparametric econometrics: the np package. *J. Statist. Softwr.*, **27**, 1–32.
Hayfield, T. and Racine, J. S. (2011) **np**: nonparametric kernel smoothing methods for mixed data types. *R Package Version 0.40-12.* (Available from `http://CRAN.R-project.org/package=np`.)
He, X. (1997) Quantile curves without crossing. *Am. Statistn*, **51**, 186–192.
Hofner, B., Hothorn, T., Kneib, T. and Schmid, M. (2011a) A framework for unbiased model selection based on boosting. *J. Computnl Graph. Statist.*, **20**, 956–971.
Hofner, B., Mayr, A., Fenske, N. and Schmid, M. (2011b) **gamboostLSS**: boosting methods for GAMLSS models. *R Package Version 1.0-3/r39.* (Available from `http://R-Forge.R-project.org/projects/gamboostlss/`.)
Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and van der Laan, M. J. (2006) Survival ensembles. *Biostatistics*, **7**, 355–373.
Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2011) **mboost**: model-based boosting. *R Package Version 2.1-1.* (Available from `http://CRAN.R-project.org/package=mboost`.)
Hothorn, T., Kneib, T. and Bühlmann, P. (2012) Conditional transformation models (extended version). *Technical Report arXiv:1201.5786v2.* Ludwig-Maximilians-Universität München, Munich. (Available from `http://arxiv.org/abs/1201.5786`.)
International Institute for Population Sciences and ORC Macro (2000) National Family Health Survey (NFHS-2), 1998–1999: India. International Institute for Population Sciences, Mumbai. (Available from `http://www.measuredhs.com/pubs`.)
Koenker, R. (2005) *Quantile Regression*. New York: Cambridge University Press.
Koenker, R. (2011) **quantreg**: quantile regression. *R Package Version 4.76.* (Available from `http://CRAN.R-project.org/package=quantreg`.)
Koenker, R. and Bassett, G. (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
Koenker, R., Ng, P. and Portnoy, S. (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–680.
Kriegler, B. and Berk, R. (2010) Small area estimation of the homeless in Los Angeles: an application of cost-sensitive stochastic gradient boosting. *Ann. Appl. Statist.*, **4**, 1234–1255.
Li, Q. and Racine, J. S. (2008) Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *J. Bus. Econ. Statist.*, **26**, 423–434.
Lu, W. and Li, L. (2008) Boosting method for nonlinear transformation models with censored survival data. *Biostatistics*, **9**, 658–667.
Lu, W. and Zhang, H. H. (2010) On estimation of partially linear transformation models. *J. Am. Statist. Ass.*, **105**, 683–691.
Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012a) Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *Appl. Statist.*, **61**, 403–427.
Mayr, A., Hothorn, T. and Fenske, N. (2012b) Prediction intervals for future BMI values of individual children—a non-parametric approach by quantile boosting. *BMC Med. Res. Methodol.*, **12**, article 6.
R Development Core Team (2011) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Ridgeway, G. (2002) Looking for lumps: boosting and bagging for density estimation. *Computnl Statist. Data Anal.*, **38**, 379–392.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Schemper, M. and Henderson, R. (2000) Predictive accuracy and explained variation in Cox regression. *Biometrics*, **56**, 249–255.

Schild, R. L., Maringa, M., Siemer, J., Meurer, B., Hart, N., Goecke, T. W., Schmid, M., Hothorn, T. and Hansmann, M. E. (2008) Weight estimation by three-dimensional ultrasound imaging in the small fetus. *Ultrasnd Obstetr. Gyn.*, **32**, 168–175.

Schmid, M. and Hothorn, T. (2008) Flexible boosting of accelerated failure time models. *BMC Bioinform.*, **9**, article 269.

Schmid, M., Hothorn, T., Maloney, K. O., Weller, D. E. and Potapov, S. (2011) Geoadditive regression modeling of stream biological condition. *Environ. Ecol. Statist.*, **18**, 709–733.

Schnabel, S. K. and Eilers, P. H. C. (2012) Simultaneous estimation of quantile curves using quantile sheets. *AStA Adv. Statist. Anal.*, to be published, doi 10.1007/s10182-012-0198-1.

Sexton, J. and Laake, P. (2012) Boosted coefficient models. *Statist. Comput.*, **22**, 867–876.

Shen, X. (1998) Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*, **85**, 165–177.

Stasinopoulos, M., Rigby, B. and Akantziliotou, C. (2011) **gamlss**: generalized additive models for location scale and shape. *R Package Version 4.1-1.* (Available from `http://CRAN.R-project.org/package=gamlss`.)

Tutz, G. and Groll, A. (2012) Likelihood-based boosting in binary and ordinal random effects models. *J. Computnl Graph. Statist.*, to be published, doi 10.1080/10618600.2012.694769.

Tutz, G. and Petry, S. (2012) Nonparametric estimation of the link function including variable selection. *Statist. Comput.*, **22**, 545–561.

Wu, C., Tian, X. and Yu, J. (2010) Nonparametric estimation for time-varying transformation models with longitudinal data. *J. Nonparam. Statist.*, **22**, 133–147.

Zeng, D., Lin, D. Y. and Yin, G. (2005) Maximum likelihood estimation for the proportional odds model with random effects. *J. Am. Statist. Ass.*, **100**, 470–483.

Zheng, S. (2012) QBoost: predicting quantiles with boosting for regression and binary classiffication. *Exprt Syst. Applic.*, **39**, 1687–1697.